Machine Translation

Hiroshi Nakagawa

(Information Technology Center; Mathematical Informatics, Graduate School of Information Science and Technology; Graduate School of Interdisciplinary Information Studies, The University of Tokyo)

Past Machine Translation

- Input sentence: "*wa-ta-si ha ri-n-go wo ta-be-ta* ‘I ate an apple.’ "
-
  - -> Morphological Analysis -> Syntax Analysis
  - noun verb noun -> subject predicate object
    - -> Semantic Analysis
    - (action = "*ta-be-ru"*, agent = "*wa-ta-si"*, target = "*ri-n-go"*, time=past)
      - Convert based on English lexicological information (convert semantic expressions at a proximity depth. <- Use a bilingual dictionary
        - (action=eat, agent=I, target=an apple, time=past)
    - Generate syntaxes and morphemes (the conversion of word order) and output translation.
      <- Use a bilingual dictionary
  - noun=I, verb(past)=ate, noun=an apple
- Output sentence: "I ate an apple."

Past Machine Translation

- It was presupposed that Japanese and English were identical in details at semantic levels.
- It was assumed that morphological, syntax, and semantic analysis were properly completed.
- However, the reality was not so simple.
  - Examples where concepts were not matched at semantic levels.
    - *Yu* -> ‘hot water’
    - *Mo-tta-i-na-i* -> ? ‘too much’, ‘no use’
    - No custom of using "personal checks" in Japan!

Bilingual Dictionary

- Japanese-> meaning
  - *Ri-n-go* -> ‘APPLE ‘
- Meaning-> English
  - ALLPE-> if bear noun or singular: apple
    if plural: apples

- At the syntax and morphological level, "an apple" is chosen as a singular form, and "apples" as a plural form.

Past Machine Translation: Example-based Translation

- Create DB to allow a parallel search for bilingual sentences.
  - E.x. "*wa-ta-si ha mi-ka-n wo ta-be-ta"* <- -> 'I ate an orange.'

- Input sentence: "*wa-ta-si ha ri-n-go wo ta-be-ta"*
  - Search for a similar Japanese sentence from bilingual DB.
    - *"wa-ta-si ha mi-ka-n wo ta-be-ta"*
      - Exchange an identical part *"mi-ka-n"* for *"ri-n-go"*.
    - Exchange *"ri-n-go"* for "an apple" based on a Japanese-English dictionary.

- Output sentence: "I ate an apple."
  - Obviously, articles, etc. were selected according to grammatical rules. Thus, a substantial amount of process utilized with existing syntax rules and morphological analysis techniques.

Previous Machine Translation: Example-based Translation

- Parallel search was an integral part. Syntax analysis could have been implemented, which would make the system closer to classical machine translation.
- Increasing collections of parallel sentences improved the quality of translation.
  - The collection activity was an automatic process.

Statistic Machine Translation (SMT)

- This method is to find counterpart translation in a target language without using linguistic knowledge. Anti linguistic theory.
- An accumulated bilingual parallel corpus.
- An aligned sentence corpus.
- The corpus is utilized to automatically align words and phrases, and extract parallel translations.
  - Sentences correspond each other. Unclear that word strings are properly in parallel.
  - A huge search space.
- Peter Brown, S. Della Pietra, V. Della Pietra, Robert Mercer, et. al. at IBM presented a paper "The Mathematics of Statistical Machine Translation: Parameter Estimation" in 1993 regarding CL. The following description is referring to this paper.

Bayes' Theorem

- Canadian Hansard: French-English Bilingual corpus
- Find a right English string:*e* for a French string:*f*.
- Given French string: *f*, find $\hat{e}=\arg\max_e \Pr(e|f)$
  - Too many candidates for *e* which correspond to *f* can be found!!

- then

$$\Pr(\boldsymbol{e}|\boldsymbol{f}) = \frac{\Pr(\boldsymbol{e})\Pr(\boldsymbol{f}|\boldsymbol{e})}{\Pr(\boldsymbol{f})}$$

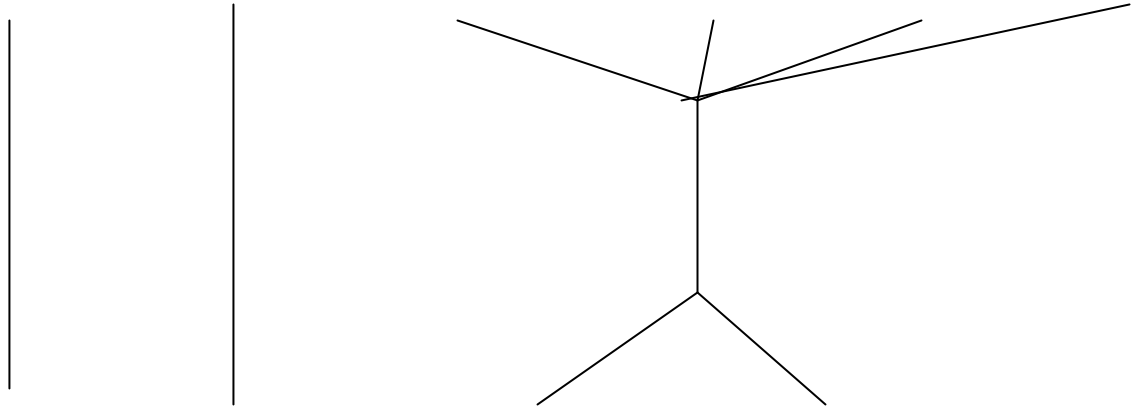
$$\Rightarrow \quad \hat{\boldsymbol{e}} = \underset{\boldsymbol{e}}{\operatorname{argmax}} \Pr(\boldsymbol{e})\Pr(\boldsymbol{f}|\boldsymbol{e})$$

Reason for Pr(*f*|*e*) x Pr(*e*) but not Pr(*e*|*f*)

- Too many candidates for *e* which correspond to *f* can be found!!
  - Parallel corpora include a small number of parallel sentences.
- The purpose of this process is to find a right English string for an unlimited number of French strings *f.*
- It is not guaranteed that the probability is high for the right English string *e* in the equation of *Pr(e|f)*.
- Obtain *Pr(e)* from separate information sources so that the issue of finding the right English is directly taken care of.

Alignment

- The$_1$ poor$_2$ don't$_3$ have$_4$ any$_5$ money$_6$



- Les$_1$ pauvres$_2$ sont$_3$ demunis$_4$

(Les pauvres sont demunis |

The(1) poor(2) don't(3,4) have(3,4) any(3,4) money(3,4))

=A(e,f)=**a**

-> *e,f* are sentences.

Description Method

- *Pr(f,a|e)* takes alignment into consideration

$$\Pr(\mathrm{f}\,|\,\mathrm{e}) = \sum_{\mathrm{a}} \Pr(\mathrm{f},\mathrm{a}\,|\,\mathrm{e})$$

$\mathrm{e} = e_1^l \equiv e_1 e_2 .... e_l$ *where l English words*

$\mathrm{f} = f_1^m = f_1 f_2 .... f_m$ *where m french words*

$\mathrm{a} = a_1^m = a_1 a_2 ..... a_{m,}$ *where* $a_j = i$

f, e express word sequence; a is alignment; f_i, e_{a_j} represent words.

$$\Pr(\mathrm{f},\mathrm{a}\,|\,\mathrm{e}) = \Pr(m\,|\,\mathrm{e}) \prod_{1}^{\mathrm{m}} \Pr(a_j\,|\,a_1^{j-1}, f_1^{j-1}, m, \mathrm{e}) \Pr(f_j\,|\,a_1^{j}, f_1^{j-1}, m, \mathrm{e})$$

- The following slides explain some methods to evaluate *Pr(f,a,|e).*

IBM Model 1

- In this model, the order of occurrence of words in English and French sentences are not considered to be inter-related. - (1)
- Translation is dependent only on individual words. - (2)

$$\varepsilon \equiv \Pr(m \mid \mathrm{e})$$

$$\Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathrm{e}) = (l+1)^{-1} \qquad -(1)$$

$$\Pr(f_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathrm{e}) = t(f_j \mid e_{a_j}) \qquad -(2)$$

$$\Pr(\mathrm{f}, \mathrm{a} \mid \mathrm{e}) = \frac{\varepsilon}{(l+1)^m} \prod_1^{\mathrm{m}} t(f_j \mid e_{a_j}) \qquad -(3)$$

f, e express word sequence; a is alignment; f_i, e_{a_j} represent words.

Model 1

- In this model, alignment a_j takes any value from 0 to m and equals 1/(l+1). Maximize $Pr(f|e)$ by Lagrange multiplier method.

$$\Pr(\mathrm{f}\,|\,\mathrm{e}) = \frac{\varepsilon}{(l+1)^m}\sum_{a_{1=0}}^{l}\cdots\sum_{a_{m=0}}^{l}\prod_{1}^{\mathrm{m}} t(f_j\,|\,e_{a_j}) \qquad -(4)$$

$$\text{constraint}: \sum_{f} t(f\,|\,e\,) = 1 \qquad f,e \text{ express either } f_j, e_{aj}.$$

$$h(t,\lambda) \equiv \frac{\varepsilon}{(l+1)^m}\sum_{a_{1=0}}^{l}\cdots\sum_{a_{m=0}}^{l}\prod_{j=1}^{\mathrm{m}} t(f_j\,|\,e_{a_j}) - \sum_{e}\lambda_e\Big(\sum_{f} t(f\,|\,e\,) - 1\Big) \qquad -(5)$$

$$0 = \frac{\partial h}{\partial t(f\,|\,e)} = \frac{\varepsilon}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\sum_{j=1}^{m}\delta(f,f_j)\delta(e,e_{a_j})t(f\,|\,e)^{-1}\prod_{k=1}^{\mathrm{m}} t(f_k\,|\,e_{a_k}) - \lambda_e \qquad -(6)$$

$$t(f \mid e) = \lambda_e^{-1} \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \prod_{k=1}^{m} t(f_k \mid e_{a_k})$$

$$= \lambda_e^{-1} \sum_{a} \Pr(\mathrm{f}, \mathrm{a} \mid \mathrm{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \qquad -(7)$$

$$c(f \mid e; \mathrm{f}, \mathrm{e}) = \sum_{a} \Pr(\mathrm{a} \mid \mathrm{f}, \mathrm{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \qquad -(8)$$

- Denote *c*(…) as the times English word *e* is translated into French word *f* in translation *(f|e)*. The second Σ represents the total number of translations of *f* and *e* in alignment *a.*

$$\Pr(\mathrm{a} \mid \mathrm{f}, \mathrm{e}) = \Pr(\mathrm{f}, \mathrm{a} \mid \mathrm{e}) / \Pr(\mathrm{f} \mid \mathrm{e}) \qquad \mathrm{f}, \mathrm{a}, \mathrm{e} \quad : \text{word string}$$

$$\rightarrow c(f \mid e; \mathrm{f}, \mathrm{e}) = \frac{\sum_{a} \Pr(\mathrm{f}, \mathrm{a} \mid \mathrm{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})}{\Pr(\mathrm{f} \mid \mathrm{e})} - (9)$$

$$\rightarrow t(f \mid e) = \lambda_e^{-1} \Pr(\mathrm{f} \mid \mathrm{e})\, c(f \mid e; \mathrm{f}, \mathrm{e}) \qquad -(10)$$

- Equation (9) $\sum_{j=1}^{m}\delta(f,f_j)\sum_{i=0}^{l}\delta(e,e_i)$ represents the number of translations of *f* and *e.* (In this equation, alignment *a* is not used.) Refer to the equation below:

- **f=** f1, f2(=f), f3, ..,f7(=f), …. fm
- e1(=e) * *
- e2
- **e=** :
- e8(=e) * *
- :
- el

- *S* as the number of translations $(f^{(s)}|e^{(s)})$ $s=1,\ldots,S$ is given as teaching data based on the corpus. Use the total number of *S* translations to rewrite equation (10). Please remember the equation below because it will be used later. Replace $\lambda_e Pr(f|e)$ in equation (10) for λ_e.

$$t(f|e)=\lambda_e^{-1}\sum_{s=1}^{S}c(f|e;f^{(s)}e^{(s)})$$

One more step to compute *t(f|e)*

- $t(f_j \mid e_{a_i})$ is a monominal expression.

$$\sum_{a_{1=0}}^{l} \cdots \sum_{a_{m=0}}^{l} \prod_{j=1}^{\mathrm{m}} t(f_j \mid e_{a_j}) = \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j \mid e_i) \qquad -(11)$$

- E.x. $t_{10}t_{20} + t_{10}t_{21} + t_{11}t_{20} + t_{11}t_{21} = (t_{10} + t_{11})(t_{20} + t_{21})$
- Therefore,

$$\Pr(\mathrm{f} \mid \mathrm{e}) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j \mid e_i) - (12)$$

- Lagrange multiplier method is again used.

$$\text{constraint}: \sum_f t(f \mid e) = 1$$

$$h(t,\lambda) \equiv \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} \sum_{c} t(f_j \mid e_i) - \sum_e \lambda_e \left(\sum_f t(f \mid e) - 1 \right) \quad -(13)$$

$$0 = \frac{\partial h}{\partial t(f \mid e)} = \frac{\varepsilon}{(l+1)^m \sum_{i=0}^{l} \sum_{c} t(f_j \mid e_i)} \prod_{j=1}^{m} \sum_{i=0}^{l} \sum_{c} t(f_j \mid e_i) \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i) - \lambda_e$$

$$-(14)$$

$$\rightarrow \lambda_e^{-1} \Pr(\mathrm{f} \mid \mathrm{e}) = \frac{\sum_{i=0}^{l} t(f_j \mid e_i)}{\sum_{j=1}^{m} \delta(f, f_j) \sum_{i=0}^{l} \delta(e, e_i)} \quad -(15)$$

$$by(10)(15) \rightarrow \qquad c(f \mid e; \mathrm{f}, \mathrm{e}) = \frac{t(f \mid e)}{t(f \mid e_0) + \ldots + t(f \mid e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=0}^{l} \delta(e, e_i) - (16)$$

Estimate *t(f|e)* with EM -1

1. Set any default value for *t(f|e).*
2. Calculate the equation below for each ($f^{(s)}$,$e^{(s)}$), *1*=<*s* =<*S*.

$$c(f|e;\mathrm{f}^{(\mathrm{s})},\mathrm{e}^{(\mathrm{s})}) = \frac{t(f|e)}{t(f|e_0)+\ldots+t(f|e_l)} \sum_{j=1}^{m} \delta(f,f_j) \sum_{i=0}^{l} \delta(e,e_i)$$

It is not only when *f* and *e* are the factors for $f^{(s)}$and $e^{(s)}$ respectively when the value of $\sum_{j=1}^{m} \delta(f,f_j) \sum_{i=0}^{l} \delta(e,e_i)$ is 0.

Estimate *t(f|e)* with EM -2

3. When $t(f \mid e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f \mid e; \mathrm{f}^{(s)} \mathrm{e}^{(s)})$ is $\sum_f$, the left-hand member becomes one.

$$\lambda_e = \sum_f \sum_{s=1}^{S} c(f \mid e; \mathrm{f}^{(s)}, \mathrm{e}^{(s)})$$

Use λ_e to estimate a new value for *t(f|e).*

$$t(f \mid e) = \frac{\sum_{s=1}^{S} c(f \mid e; \mathrm{f}^{(s)}, \mathrm{e}^{(s)})}{\sum_f \sum_{s=1}^{S} c(f \mid e; \mathrm{f}^{(s)}, \mathrm{e}^{(s)})}$$

4. Repeat 2 & 3 until *t(f|e)* is converged.

Model 2

- Alignment depends on locations.

$$a(a_j \mid j, m, l) \equiv \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, l)$$

is dependent on *j*, a_j, *m*, and *l*.

$$\sum_{i=0}^{l} a(i \mid j, m, l) = 1$$

then $$\Pr(\mathrm{f} \mid \mathrm{e}) = \varepsilon \sum_{a_1=0}^{l} \ldots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j \mid e_{a_j}) a(a_j \mid j, m, l)$$

Lagrange

$$h(t, a, \lambda, \mu)$$

$$= \Pr(\mathrm{f} \mid \mathrm{e}) - \sum_{\mathrm{e}} \lambda_{\mathrm{e}} \left(\sum_{f} t(f \mid e) - 1\right) - \sum_{j} \mu_{jml} \left(\sum_{i} a(i \mid j, m, l) - 1\right)$$

Use Lagrange multiplier method to obtain differentiated *h.*

$$t(f \mid e) = \lambda_e^{-1} \sum_a \Pr(f, a \mid e) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_j)$$

$$c(f \mid e; \mathrm{f}, \mathrm{e}) = \sum_a \Pr(\mathrm{a} \mid \mathrm{f}, \mathrm{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})$$

$$t(f \mid e) = \lambda_e^{-1} \Pr(f \mid e)\, c(f \mid e; \mathrm{f}, \mathrm{e})$$

$$t(f \mid e) = \lambda_e^{-1} \Pr(f \mid e) \sum_{s=1}^{S} c(f \mid e; \mathrm{f}^{(\mathrm{s})}\, \mathrm{e}^{(\mathrm{s})})$$

Same procedures until the line above. Semantic analysis is performed…

$$c(i \mid j, m, l; \mathrm{f}, \mathrm{e}) = \sum_a \Pr(\mathrm{a} \mid \mathrm{e}, \mathrm{f})\, \delta(i, a_j)$$

$$a(i \mid j, m, l) = \mu_{jml}^{-1} \sum_{s=1}^{S} c(i \mid j, m, l; \mathrm{f}^{(\mathrm{s})}, \mathrm{e}^{(\mathrm{s})})$$

Note $\Pr(a \mid f, e) \Pr(f \mid e) = \Pr(a, f \mid e)$

$\mu_{jml} \Pr(f \mid e) \rightarrow \mu_{jml}$

Apply the same method as Model 1.

- *a(i|j,m,l)* is $(l+1)^{-1}$ in Model 1 while it is variables in Model 2.

$$\mathbf{Pr}(\mathrm{f}\,|\,\mathrm{e}) = \varepsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j \,|\, e_i)\, a(i \,|\, j,m,l)$$

$$c(f \,|\, e; \mathrm{f}, \mathrm{e}) = \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(f \,|\, e)\, a(i \,|\, j,m,l)\, \delta(f, f_j)\, \delta(e, e_i)}{t(f \,|\, e_0)\, a(0 \,|\, j,m,l) + \ldots + t(f \,|\, e_l)\, a(l \,|\, j,m,l)}$$

$$c(i \,|\, j,m,l; \mathrm{f}, \mathrm{e}) = \frac{t(f_j \,|\, e_i)\, a(i \,|\, j,m,l)}{t(f_j \,|\, e_0)\, a(0 \,|\, j,m,l) + \ldots + t(f_j \,|\, e_l)\, a(l \,|\, j,m,l)}$$

- For the rest, find *t(f|e)* with EM algorithm in the same way.
- Use the result of Model 1 for the default value.

Model 3

- One word is translated into *n* words: not => ne … pas
- Consider the case of *n=0* (not translated). Japanese does not have articles.
- The locations of corresponding words are twisted.
  - Difference in word order between Japanese and English.
- Model to respond to these conditions.

- Reproduction probability $n(\phi|e)$: Probability that English word *e* is linked to ϕ number of French words.
- Translation probability $t(f|e)$: Probability that English word *e* is translated to French word *f*.
- Distortion probability $d(j|i,m,l)$: Probability that the length of English sentence *l,* the length of French sentence *m*, and the location of English word *i* are connected to the location of French word *j.*
- The number of reproduction of empty English word= ϕ_0

- Φ_0 number of words are inserted at the probability of p_1 after French words are generated based non-empty English words:

$$\Pr(\ \phi_0 \mid \phi_1^{\,l}, e\ ) = \binom{\phi_1 + .. + \phi_l}{\phi_0}\, p_0^{\,\phi_1 + .. + \phi_l - \phi_0}\ p_1^{\,\phi_0}$$

$$p_0 + p_1 = 1$$

$$\sum_{i=0}^{m} \phi_i = m$$

After the following preparation:

$$\Pr(\mathrm{f} \mid \mathrm{e}) = \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \Pr(\mathrm{f}, \mathrm{a} \mid \mathrm{e})$$

$$= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! n(\phi_i \mid e_i) \times$$

$$\prod_{j=1}^{m} t(f_j \mid e_{a_j}) d(j \mid a_j, m, l) \qquad -(32)$$

$\Phi_i!$ represents the count of word strings whose order in the word string of Φ_i number generated from e_i Is changed.

- Maximize *Pr(e|f)* in equation (32) by Lagrange multiplier method on condition that the addition of *n, t, d,* and $p_{0,1}$ totals 1.
- However, unlike model 1 and 2, it must be directly calculated because addition and multiplication are not interchangeable.
- There are many combinations. Viterbi algorithm is used to perform a rough calculation.